Signaux stationnaires sur graphe : étude d'un cas réel

Benjamin GIRAULT[1,3], Paulo GONÇALVES[2,3], Éric FLEURY[1,2,3]

[1]École Normale Supérieure de Lyon, Université de Lyon

[2]Inria Rhône-Alpes

[3]LIP (UMR 5668 : CNRS – ENS Lyon – UCB Lyon 1 – Inria) – Labex Milyon – IXXI

benjamin.girault@ens-lyon.fr, paulo.goncalves@ens-lyon.fr, eric.fleury@ens-lyon.fr

Résumé – Sur un jeu de données géographiques réelles, nous appliquons la caractérisation de la propriété de stationnarité d'un signal sur graphe via l'analyse de ses coefficients spectraux. Nous identifions différentes sources possibles de non-stationnarité et isolons l'influence qu'a le graphe sous-jacent sur la cohérence structurelle des données.

Abstract – Based on a real geographical dataset, we apply the stationarity characterisation of a graph signal, through the analysis of its spectral decomposition. In the course, we identify possible sources of non-stationarity and we elaborate on the impact of the graph used to model the structural coherence of the data.

1 Introduction

Le développement de moyens technologiques toujours plus rapides et précis produit un déluge de données de plus en plus complexes, mais également de plus en plus structurées. On citera par exemple les réseaux sociaux, les réseaux informatiques entre ordinateurs ou systèmes intelligents, les réseaux biologiques, ou les mesures géolocalisées.

Le domaine émergeant du traitement du signal sur graphe entend répondre à la problématique de l'étude de ces données en exploitant les forces des outils du traitement du signal classique pour les étendre aux signaux portés par des graphes. Les avancées récentes du domaine incluent le filtrage [10], la TFCT [11], les ondelettes [7], l'EMD [14], ou le principe d'incertitude [1].

Nous avons récemment introduit la notion de signal stationnaire sur graphe [5] grâce à l'opérateur *graph translation* [6]. Dans cette communication nous détaillons l'étude empirique de la propriété de stationnarité d'un signal sur graphe par l'utilisation de données météorologiques de terrain. Ce faisant, nous détaillons les sources de non-stationnarité observées dans ces données, en mettant en évidence les rôles respectifs de la statistique des mesures et du graphe qui supporte et structure ces mesures.

2 Signaux stationnaires sur graphe

2.1 Traitement du signal sur graphe

Soit $\mathcal{G} = (V, E)$ un *graphe* avec V ses *nœuds* et $E \subseteq V \times V$ ses *arêtes*. On supposera les arêtes non orientées, *i.e.* si ij est une arête, ji l'est également. Soit A la *matrice d'adjacence* du graphe telle que a_{ij} est le *poids* de l'arête ij, ou 0 si $ij \notin E$. Soit $L = D - A$ le *Laplacien* du graphe, avec D la matrice diagonale des *degrés* $d_{ii} = \sum_j a_{ij} = d_i$. Les arêtes n'ayant pas d'orientation, L est une matrice semi-définie positive [3].

On note $X : V \to \mathbb{R}$ ou $\mathbb{C}$ un *signal sur le graphe* $\mathcal{G}$. On définit la *matrice de Fourier* sur le graphe $\mathcal{G}$ par la décomposition spectrale du Laplacien $L\chi_l = \lambda_l \chi_l$ telle que $L = F^* \Lambda F$, avec Λ la matrice diagonale des valeurs propres de L [12]. La *transformée de Fourier* de X est alors la multiplication matricielle $\widehat{X} = FX$, où X est représenté sous la forme d'un vecteur colonne.

On remarquera que contrairement aux modes de Fourier classiques, les modes de Fourier χ_l d'un graphe n'ont pas tous un support s'étendant sur l'ensemble des nœuds, mais peuvent être fortement localisés [13]. De tels modes sont illustrés sur la Fig. 3. Cette possible localisation des modes de Fourier ne change pas les définitions de la partie suivante, mais joueront un rôle dans l'interprétation des résultats.

2.2 Stationnarité

On s'intéresse dans cette communication à l'étude de signaux sur graphe aléatoires. On notera $\mathbf{X}$ un tel signal. On rappelle dans cette partie l'opérateur *graph translation* [6] et les propriétés de stationnarité [5] introduites précédemment.

On définit les *fréquences de graphe* à partir des valeurs propres de L par :

$$\omega_l = \pi \sqrt{\lambda_l / \rho_{\mathcal{G}}}, \tag{1}$$

où $\rho_{\mathcal{G}}$ est une borne supérieure des valeurs propres de L [4] :

$$\rho_{\mathcal{G}} = \max_{i \in V} \sqrt{2 d_i (d_i + \bar{d}_i)} \quad \text{avec} \quad \bar{d}_i = \frac{\sum_{ij \in E} w_{ij} d_j}{d_i}.$$

Les fréquences de graphe appartiennent à l'intervalle $[0, \pi]$, et une valeur faible de ω_l correspond à un mode de Fourier χ_l de

basse fréquence. On définit ensuite l'opérateur *graph translation* $T_{\mathcal{G}}$ par :

$$T_{\mathcal{G}} = \exp\left(\imath\pi\sqrt{L/\rho_{\mathcal{G}}}\right). \quad (2)$$

Le translaté du signal X est alors $T_{\mathcal{G}}X$, et on a $T_{\mathcal{G}}\chi_l = e^{\imath\omega_l}\chi_l$.

De la même manière qu'un signal temporel est SSS s'il est invariant en loi par l'opérateur de translation en temps, on définit la *stationnarité forte* (SSS) comme l'invariance en loi d'un signal aléatoire sur graphe par l'opérateur *graph translation* :

$$\mathbf{X} \overset{\mathrm{d}}{=} T_{\mathcal{G}}\mathbf{X}. \quad (3)$$

SSS étant une propriété très contraignante sur les signaux, on définit également la *stationnarité faible* (WSS) comme l'invariance des deux seuls premiers moments :

$$\mathbb{E}[\mathbf{X}] = \mathbb{E}[T_{\mathcal{G}}\mathbf{X}] \quad (4)$$

$$\mathbb{E}[\mathbf{X}\mathbf{X}^*] = \mathbb{E}[(T_{\mathcal{G}}\mathbf{X})(T_{\mathcal{G}}\mathbf{X})^*]. \quad (5)$$

On note $\mu_{\mathbf{X}} = \mathbb{E}[\mathbf{X}]$ le premier moment et $R_{\mathbf{X}} = \mathbb{E}[\mathbf{X}\mathbf{X}^*]$ le second moment de $\mathbf{X}$. $R_{\mathbf{X}}$ est alors la *matrice de corrélation* du signal X. On note $S_{\mathbf{X}} = \mathbb{E}[\widehat{\mathbf{X}}\widehat{\mathbf{X}}^*]$ la *matrice de corrélation des composantes fréquentielles* de $\mathbf{X}$. On obtient alors $S_{\mathbf{X}} = FR_{\mathbf{X}}F^*$, rappelant le théorème de Wiener-Khintchine [9]. De plus, les signaux WSS ont une caractérisation spectrale que nous utiliserons et étudieront dans cette communication [5] :

Proposition 1 (Caractérisation spectrale)**.** *Un signal* $\mathbf{X}$ *est faiblement stationnaire si et seulement si :*

1. $\mu_{\mathbf{X}} \propto \chi_0$
2. *si* $\omega_l \neq \omega_k$, *alors* $(S_{\mathbf{X}})_{lk} = 0$

En d'autre termes, un signal est WSS si son premier moment est constant (χ_0 est un vecteur constant) et si ses composantes spectrales sont décorrélées. De plus, si toutes les fréquences de graphe sont différentes, la condition 2. est équivalente à $S_{\mathbf{X}}$ diagonale.

Le lecteur trouvera les détails de ces notions dans les communications [5, 6].

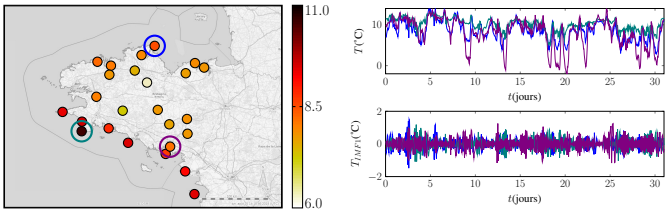
3 Application sur des données réelles

Nous étudions maintenant la stationnarité de données de terrain portées par un graphe afin de mettre en pratique les notions théoriques introduites dans [5, 6], mais également dans le but de décrire un mode opératoire pour l'utilisation des estimateurs empiriques sur lesquels repose la caractérisation de la stationnarité.

3.1 Données météorologiques

Le jeu de données étudié a été publié par Météo France [1] et concerne les données horaires relevées par les stations météorologiques de Bretagne sur le mois de Janvier 2014. Cela représente 744 mesures horaires par station, sur 25 stations, de la force du vent, de la température et des précipitations.

1. https ://www.data.gouv.fr/



(a) Température moyenne. (b) Variations temporelles.

Fig. 1 – Variations spatiales et temporelles de la température en Bretagne durant le mois de Janvier 2014. (a) montre la température moyenne en chacune des stations, et (b) les variations temporelles de trois stations des données brutes (haut) et de la première IMF (bas).

Chacune de ces stations est géolocalisée, permettant ainsi de créer un graphe basé sur les distances géographiques. La Fig. 1(a) montre ces stations sur un fond de carte. Pour l'étude de ces données, nous utilisons le graphe complet pondéré par un noyau Gaussien de la distance : $a_{ij} = \exp(-\kappa\,\mathrm{d}(i,j)^2)$ avec $\kappa = 10^{-9}$ et $\mathrm{d}(i,j)$ la distance en kilomètres entre les stations i et j. En particulier, le poids maximal est de 0.87 et si $\mathrm{d}(i,j) > 96$km, alors $a_{ij} < 10^{-4}$. Cette pondération possède de bonnes propriétés pour étudier ce type de données [2].

La question que nous étudions est alors celle de la caractérisation des propriétés de (non-)stationnarité sur ce jeu de données.

3.2 Pré-traitement

La stationnarité de la Proposition 1 étant caractérisée par des moments statistiques, nous devons les estimer via des estimateurs empiriques. S'agissant du premier moment, on peut immédiatement conclure que les données brutes ne sont pas stationnaires. En effet, la moyenne temporelle n'est pas constante sur les nœuds, comme illustré sur la Fig. 1(a) pour la température, donc ce premier moment n'est pas colinéaire à χ_0. Le même phénomène est observable sur le vent.

On centre donc chacune des séries temporelles pour étudier la variabilité des mesures autour de ces moyennes. Malheureusement, les instantanés de ces données centrées ne sont pas indépendants puisque chacune de ces séries temporelles possède une tendance temporelle forte composée entre autres d'une déviation et de variations saisonnières. Ces tendances apparaissent nettement sur le graphique supérieur de la Fig. 1(b). Or, pour pouvoir utiliser des estimateurs empiriques (via l'hypothèse d'ergodicité) des matrices R et S, il faut s'assurer que les séries temporelles de chacune des stations soient stationnaires au cours du temps.

Pour supprimer cette tendance, on utilise la *décomposition modale empirique* (EMD) [8]. Cette décomposition sépare une série temporelle en un ensemble de modes de variations rapides (première *Intrinsic Mode Function* (IMF)) à lente (dernière IMF). La première IMF ne comporte alors plus de tendance temporelle. Le résultat est illustré sur le graphique inférieur de la Fig. 1(b).

Cette opération produit de bons résultats sur le vent et la température. En revanche, l'EMD atteint ses limites dans le

cas des précipitations. En effet, ces données possèdent de forts phénomènes de saturation en cas d'absence de précipitations qui rendent l'utilisation de l'EMD inadéquate. Dans le reste de cette communication, nous étudions les IMF1 du vent et de la température.

3.3 Étude de la stationnarité

La stationnarité se caractérisant par la diagonalité de S, il est plus aisé de travailler avec la matrice des coefficients de corrélation spectrale afin d'apprécier cette diagonalité de S. Soit $\mathbf{x}$, $\mathbf{y}$ deux variables aléatoires. Leur coefficient de corrélation est :

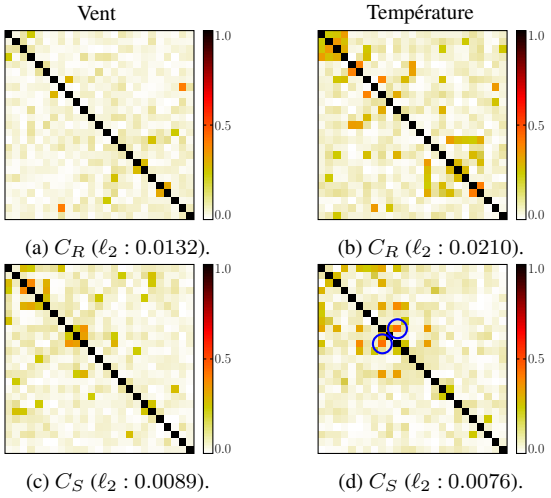
$$\mathrm{c}(\mathbf{x}, \mathbf{y}) = \mathrm{cov}(\mathbf{x}, \mathbf{y})/(\sigma_{\mathbf{x}}\sigma_{\mathbf{y}})$$

avec $\mathrm{cov}(\mathbf{x}, \mathbf{y})$ leur covariance, et $\sigma_{\mathbf{x}}$ l'écart-type de $\mathbf{x}$. On notera C_S la matrice des coefficients de corrélation associée à S et telle que $(C_S)_{ij} = \mathrm{c}(\widehat{\mathbf{X}}_i, \widehat{\mathbf{X}}_j)$, et C_R celle associée à R. Puisque nos IMFs 1 sont centrées, elles sont de moyenne nulle et leurs composantes spectrales de moyenne nulle également. Dès lors R ou S diagonale équivaut à C_R ou C_S diagonale.

Les IMF1 des données de vent et de température produisent les matrices empiriques C_R et C_S montrées sur la Fig. 2. Nous remarquons sur les Fig. 2(a) et 2(b) que les matrices C_R ne sont pas diagonales, illustrant ainsi les corrélations spatiales des données, d'autant plus prononcées pour la température.

De plus, il apparaît sur les Fig. 2(c) et 2(d) que les matrices C_S ne sont pas non plus diagonales. On remarque que le plus fort coefficient de corrélation spectrale de la température correspond en fait à deux modes de Fourier co-localisés dans la partie sud-ouest, suggérant une corrélation du fait de la structure et non des données. Cette corrélation est entourée en bleu sur la Fig. 2(d) et les modes en question illustrés sur la Fig. 3. Naturellement, cette corrélation structurelle n'exclue pas que les données soient également intrinsèquement non-stationnaires.

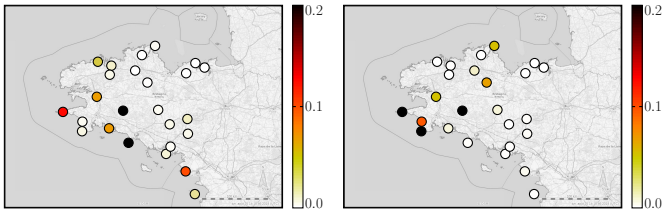
Les matrices C_S n'étant pas diagonales, ces signaux ne sont



(a) C_R (ℓ_2 : 0.0132). (b) C_R (ℓ_2 : 0.0210).

(c) C_S (ℓ_2 : 0.0089). (d) C_S (ℓ_2 : 0.0076).

Fig. 2 – Matrices des coefficients de corrélation dans les domaines des nœuds (C_R) et des composantes spectrales (C_S) pour les données de vitesse du vent et de température.



(a) Mode 9. (b) Mode 11.

Fig. 3 – Localisation de l'énergie de deux modes de Fourier du graphe des stations pondéré par un noyau Gaussien.

donc pas formellement stationnaires. Dans la suite, nous montrons comment comprendre cette non-stationnarité et séparer les corrélations spatiales dues aux mesures de celles dues à la structure.

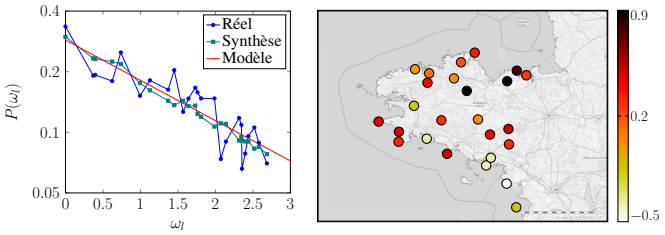
3.4 Interprétations

Nous illustrons tout d'abord, l'influence de la structure géographique des données sur les corrélations spatiales. Considérons les données de températures, et les matrices des Fig. 2(b) et 2(d). La diagonale de la matrice S associée contient les *densités spectrales de puissance* (PSD) empiriques de ces données. On observe alors que la PSD suit une loi de type loi de puissance tracée en rouge sur la Fig. 4(a) avec la PSD en bleu.

Afin d'étudier l'influence des corrélations spectrales sur les corrélations spatiales, nous utilisons ce modèle pour synthétiser des réalisations du signal stationnaire sur graphe de moyenne nulle et de matrice de corrélation S prescrite par la loi précédente. On utilise des distributions Gaussiennes sur les nœuds telles qu'observées dans les données.

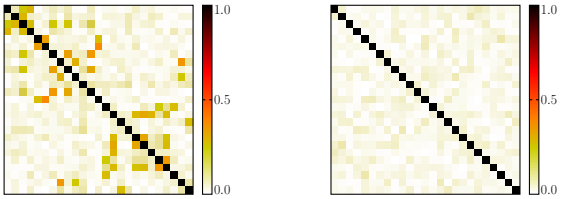
L'hypothèse Gaussienne rend la génération de ces réalisations aisée en utilisant la décomposition de Cholesky $S = LL^*$, avec L triangulaire inférieure. Soit Y une réalisation de N variables Gaussiennes centrées réduites indépendantes. On pose $\widehat{X} = LY$. La matrice de corrélation spectrale de $\mathbf{X}$ est alors S.

La Fig. 5 montre les matrices empiriques C_R et C_S de ces signaux synthétiques. Conformément au modèle, la matrice C_S est bien diagonale. Pour quantifier les différences entre ces matrices et celles des données réelles, nous proposons d'utiliser la norme ℓ_2 des coefficients significatifs (de p-valeur inférieure à 0.1) non diagonaux des matrices, divisée par le nombre de ces



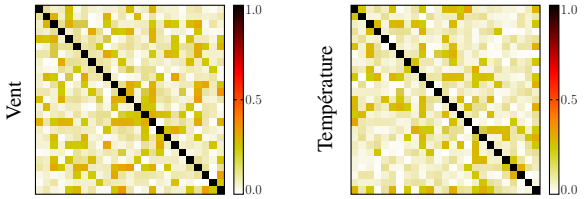
(a) PSDs. (b) Température.

Fig. 4 – Données synthétiques issues d'un modèle de PSD en $1/\omega_l^\alpha$ obtenu par régression linéaire de la PSD empirique de la température. (a) montre la PSD des données réelles superposée au modèle et à la PSD empirique obtenue avec 744 réalisations du modèle. (b) montre une de ces réalisations.



(a) C_R (ℓ_2 : 0.0224). (b) C_S (ℓ_2 : 0.0062).

Fig. 5 – Matrices des coefficients de corrélation de signaux stationnaires de synthèse de PSD prescrite en $1/\omega_l^\alpha$, avec autant de réalisations que pour la Fig. 2.



(a) C_S (ℓ_2 : 0.0190). (b) C_S (ℓ_2 : 0.0137).

Fig. 6 – Matrices des coefficients de corrélation des composantes spectrales des données de vitesse du vent et température en utilisant un graphe cyclique reliant les stations.

coefficients. On observe alors une valeur inférieure de 20% de ce critère pour les données synthétiques : le modèle ne colle pas parfaitement aux données.

Cependant, les matrices C_R des données et du modèle présentent de fortes similarités. Plus précisément, notre critère ℓ_2 appliqué à la différence des matrices C_R est de 0.0065, valeur faible devant les valeurs de ℓ_2 pour chacune des matrices. Ainsi nous pouvons expliquer nombre de corrélations spatiales et la PSD par la structure du graphe qui porte les mesures.

Réciproquement, supprimons maintenant la cohérence structurelle du graphe et étudions ce qu'il advient de la propriété de non-stationnarité. Pour cela, nous ré-organisons les stations de manière à ce qu'elles forment un graphe cyclique, tel que l'ordre de ces stations est arbitraire. Les arêtes de ce graphe n'ont alors plus aucun lien avec l'organisation géographique des stations. Nous étudions maintenant l'impact de cette réorganisation sur la stationnarité. La Fig. 6 montre les matrices C_S des données sur ce graphe cyclique. On observe une augmentation d'un facteur 2 de notre critère ℓ_2 par rapport au graphe naturel. Ces corrélations sont également réparties plus équitablement au sein de la matrice.

De plus, les corrélations spatiales ne ressemblent en rien à celles des données originales, si bien que l'organisation cyclique des stations ne permet pas d'interpréter les corrélations spatiales observées. Un tel graphe n'explique donc pas la non-stationnarité par sa structure. Dès lors, la suppression de l'organisation des liens entre les mesures fait apparaître une non-stationnarité plus prononcée comparativement à l'étude des données avec leur organisation réelle.

4 Conclusion

La stationnarité des signaux temporels est un problème difficile mais incontournable et qui soulève encore de nombreuses questions. Dans le cas des signaux sur graphe, cette difficulté est encore accentuée par le graphe qui supporte et structure les données et qui joue un rôle déterminant sur la notion même de stationnarité des données et sur sa caractérisation. Le socle théorique introduit dans [5] permet cependant d'étudier de tels signaux via une caractérisation spectrale simple, qui nous a permis, sur l'exemple réel traité dans cette communication, d'illustrer les structures corrélatives des signaux stationnaires et de mettre en évidence l'influence des différentes composantes d'un signal sur graphe sur sa (non-)stationnarité. La problématique qui se pose maintenant est celle d'un test statistique de la stationnarité sur une seule réalisation, à l'instar des signaux temporels.

Références

[1] A. Agaskar and Y. M. Lu. A Spectral Graph Uncertainty Principle. *IEEE Trans. on Inf. Th.*, 59(7) :4338–4356, 2013.

[2] M. Belkin and P. Niyogi. Towards a theoretical foundation for Laplacian-based manifold methods. *Journal of Computer and System Sciences*, 74(8) :1289–1308, 2008.

[3] F. R. K. Chung. Lectures on spectral graph theory. *CBMS Lectures, Fresno*, 1996.

[4] K. C. Das. Extremal graph characterization from the bounds of the spectral radius of weighted graphs. *Applied Mathematics and Computation*, 217(18) :7420–7426, 2011.

[5] B. Girault. Stationary Graph Signals using an Isometric Graph Translation. Soumis à Eusipco 2015.

[6] B. Girault, P. Gonçalves, and E. Fleury. Translation on Graphs : an Isometric Shift Operator. En préparation pour soumission à IEEE Signal Processing Letters.

[7] D. K. Hammond, P. Vandergheynst, and R. Gribonval. Wavelets on graphs via spectral graph theory. *Applied and Computational Harmonic Analysis*, 30(2) :129–150, 2011.

[8] N. E. Huang et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *The Roy. Soc. of Lon. Proc. Series A. Math., Phys. and Eng. Sc.*, 454(1971) :903–995, 1998.

[9] A. Papoulis. *Probability, random variables, and stochastic processes*. McGraw-Hill, 3rd edition, 1991.

[10] A. Sandryhaila and J. M. F. Moura. Discrete Signal Processing on Graphs. *IEEE Trans. on Sig. Proc.*, 61(7) :1644–1656, 2013.

[11] D. Shuman, B. Ricaud, and P. Vandergheynst. A windowed graph Fourier transform. In *Statistical Signal Processing Workshop (SSP), 2012 IEEE*, pages 133–136. IEEE, 2012.

[12] D. I. Shuman et al. The Emerging Field of Signal Processing on Graphs : Extending High-Dimensional Data Analysis to Networks and Other Irregular Domains. *IEEE SPMag.*, 30(3) :83–98, 2013.

[13] D. I. Shuman, B. Ricaud, and P. Vandergheynst. Vertex-frequency analysis on graphs . *App. and Comp. Harm. Ana.*, 2015. In press.

[14] N. Tremblay, P. Borgnat, and P. Flandrin. Graph empirical mode decomposition. In *Sig. Proc. Conf. (EUSIPCO), 2014 Proc. of the 22nd Eur.*, pages 2350–2354. IEEE, 2014.